

# Cervical Cancer Risk Factors Data: Implementing Data Science Pipelines for Analysis and ML Classification

Ioannis Stathakis

School of Science and Engineering  
University of Nicosia  
Nicosia, Cyprus

[jstatdata@gmail.com](mailto:jstatdata@gmail.com)

## ABSTRACT

Cervical cancer is one of the most common and preventable cancers affecting women worldwide, with early diagnosis shown to be key to successful intervention. Routine screening procedures such as biopsies, cytology, and visual inspections (for example Schiller and Hinselmann tests) are widely used to detect pre-cancerous changes (1-3). This study applies machine learning techniques to predict four diagnostic outcomes - Biopsy, Schiller, Hinselmann, and Cytology - based on patient medical and behavioral data. Using a cleaned and curated dataset, the project involved full-cycle data science methodology including data cleaning/preprocessing, exploratory data analysis (EDA), statistical testing, feature selection, and model building. Class imbalance was addressed with the SMOTE oversampling algorithm (4), and models were evaluated using appropriate standard metrics such as Accuracy, Precision, Recall, F1-Score and ROC-AUC. Logistic Regression emerged as the most consistently reliable classifier, while Random Forest was better suited for more imbalanced targets. The results suggest that statistical modeling can offer meaningful support in cervical cancer screening strategies.

## KEYWORDS

Cervical Cancer, Biopsy, Schiller, Hinselmann, Cytology, Machine Learning, Logistic Regression, Random Forest, XGBoost

## 1. Introduction

Cervical cancer remains a global public health challenge, ranking as the fourth most common cancer among women and causing hundreds of thousands of deaths annually. Most cases are preventable through regular screening and early treatment. Diagnostic tests such as cervical cytology (Pap smear), biopsies, and visual inspections like the Schiller and Hinselmann tests are commonly used in clinical screening. However, in many healthcare settings - especially those with limited resources - there is a need to optimize screening strategies and prioritize high-risk patients.

With increasing availability of patient medical records and risk factor data, machine learning (ML) offers an opportunity to

improve early detection by identifying patterns and predictive signals that may not be obvious through conventional analysis, and various projects have already been completed in this direction (5). This project investigates the use of ML classifiers to predict four diagnostic outcomes: Biopsy, Schiller, Hinselmann, and Cytology, using a dataset composed of clinical, behavioral, and reproductive health features.

The report outlines a complete ML workflow—beginning with data cleaning, proceeding through exploratory analysis and statistical testing, and culminating in classifier development and evaluation. Emphasis is placed on handling class imbalance, selecting relevant features, and interpreting model performance with respect to clinical applicability.

## 2. Data Cleaning Phase

The dataset contained a mix of continuous, categorical and binary features. Several preprocessing steps were applied:

- Conversion of object-type columns to numeric types
- Imputation of missing values using median/mode strategies appropriately
- Removal of sparse or low-information columns
- Detection and retention of medically relevant outliers
- Removal of duplicate rows

The cleaned dataset was saved and served as the foundation of the exploratory and modeling phases.

### 2.1. Initial Dataset Overview

The dataset (6) was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients (rows), spanning 36 columns representing various features including the four modeling target variables: Biopsy, Schiller, Hinselmann and Cytology, on which the modeling phase was focused. Several patients decided not to answer some of the questions because of privacy concerns (missing values). The raw dataset consisted of patient records with a mix of continuous, discrete, and binary features, including self-reported health information and clinical history related to cervical cancer risk factors.

## 2.2. Data Type Formatting

Initial inspection revealed several columns encoded as object types despite containing numeric values, which were systematically converted to appropriate numerical formats. Integer formatting was enforced on all binary features to ensure consistency in downstream analysis.

## 2.3. Handling Missing Values

Missing data was prevalent across 26 variables, with some STD-related columns exhibiting high sparsity (>90% NaN or extreme levels of subclass imbalance). These were considered low-informative and subsequently dropped. For the remaining features, imputation strategies were chosen based on distribution characteristics: median imputation for skewed continuous features (e.g., "Number of Pregnancies") and mode imputation for categorical or binary variables (e.g., "Smokes", "Hormonal Contraceptives"). Variables with extremely unbalanced distributions and high missingness were excluded from the modeling phase.

## 2.4. Outliers and Duplicates

Outlier detection using boxplots revealed extreme values in some clinical features, which also influenced the skewness of variable distributions. However, it was decided that outliers should be subsequently retained, due to their potential medical significance. Finally, 28 exact duplicate records were identified and removed to avoid bias in model training. The final shape of the dataset was 830 rows and 24 feature columns.

## 3. Exploratory Data Analysis (EDA)

The goal of this phase is to explore data and classes/subclasses distributions, class imbalance, correlations and relationships between features and each of the four binary target variables.

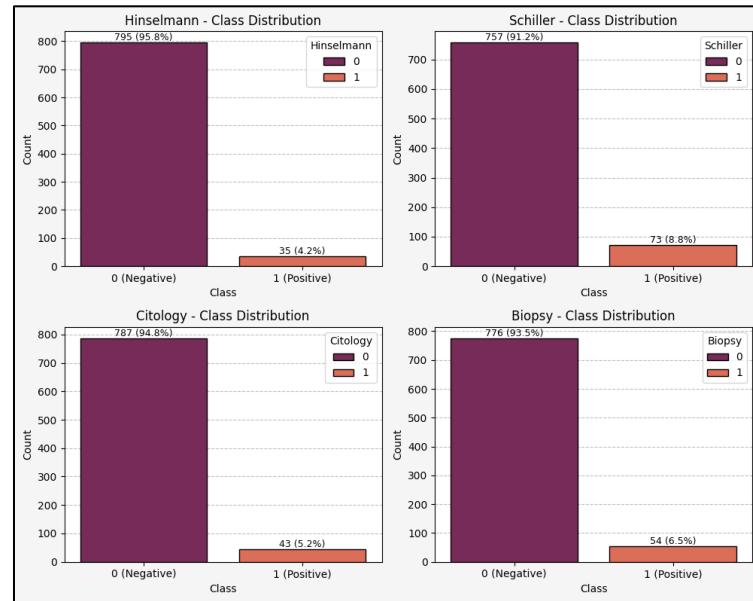
### 3.1. Target Variable Distributions

All targets were found to be significantly imbalanced, with positive subclasses representing less than 10% of the data in all cases (*Figure 1*). This was a clear indication that resampling algorithms were essential for the modeling phase, along with models which are robust to class imbalance (such as Random Forests). As will be explained in the modeling phase, the algorithm of choice ended up being SMOTE (Synthetic Minority Oversampling Technique), along with stratified train/test splits.

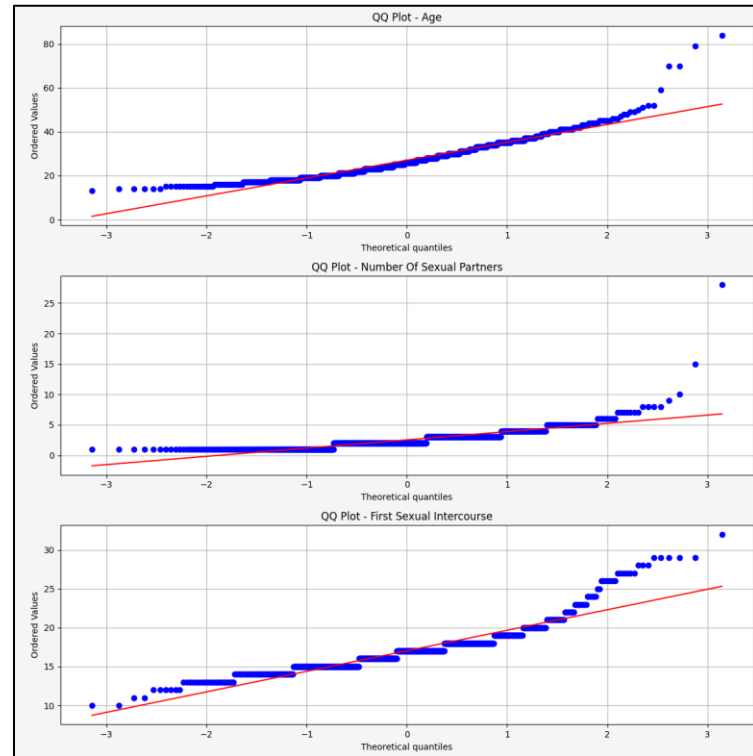
### 3.2. Normality Checks / Feature-Target Relationships

Statistical normality of the features was tested for all continuous variables, to determine the appropriate statistical tests that were to be used further down in the analysis. Normality tests employed were the Shapiro-Wilk and Anderson-Darling tests, both of which indicated non-normal distributions across all features. Complementary visual normality tests (QQ-plots) were also employed, which confirmed the aforementioned results of the statistical tests, and provided a more concise visual indication of the non-normality of the features (*Figure 2*). As a result, non-parametric alternatives were adopted for group comparisons.

Features were grouped into three different groups depending on their value characteristics, namely (a) continuous or high cardinality discrete values, (b) low cardinality discrete values, and (c) binary values. Feature distributions per target subclass were subsequently analyzed for each respective group, using kernel density estimation (KDE), violin plots, and boxplots for group (a), and stacked bar plots for groups (b) and (c), allowing visualization of class separation (*Figure 3*).



**Figure 1:** Count plots showing the subclass distribution of each binary target variable. As can be seen, there is a notable class imbalance among all targets.



**Figure 2:** Example QQ-plots, depicting the non-normality of the distribution of the features "Age", "Number of Sexual Partners", and "First Sexual Intercourse". The plots confirm the results obtained by the Shapiro-Wilk and Anderson-Darling statistical normality tests.

The strongly confirmed non-normality of the dataset features dictated the use of non-parametric tests to appropriately test for statistically significant differences between the distributions per target subclass. We opted for the Mann-Whitney U test for the features belonging to group (a), and the Chi-Squared Test for Independence for the features belonging to groups (b) and (c). Results were varied, with some statistically significant variances being observed, but also with lack of concrete statistical evidence for different variances in many features per target subclass, for all targets.

### 3.3. Correlation and Multicollinearity

Correlation heatmaps using Spearman rank correlation (as an alternative to Pearson correlation, due to non-normality) were generated to identify potential associations between features and targets. Additionally, multicollinearity was assessed using Variance Inflation Factor (VIF), revealing several highly correlated binary variables (*Table 1*). These were retained but flagged for future sensitivity analysis during model development. The EDA phase concluded with the identification of a reduced, statistically relevant feature set for each target variable.

Feature	VIF
Stds: Condylomatosis	49.275638
Stds: Vulvo-Perineal Condylomatosis	42.900193
Stds (Number)	33.582318
Stds	10.357926
Stds: Number of Diagnosis	8.179179
Dx: Cancer	6.842370
Dx	6.416449
Dx: HPV	4.912615
Smokes (Years)	3.621837
Dx: CIN	3.457821
IUD	2.503962
IUD (Years)	2.364346
Smokes	2.279434
Smokes (Packs/Year)	2.229411
Age	2.122842
Number of Pregnancies	1.565776
First Sexual Intercourse	1.402103
Hormonal Contraceptives (Years)	1.350521
Hormonal Contraceptives	1.220568
Number of Sexual Partners	1.108684

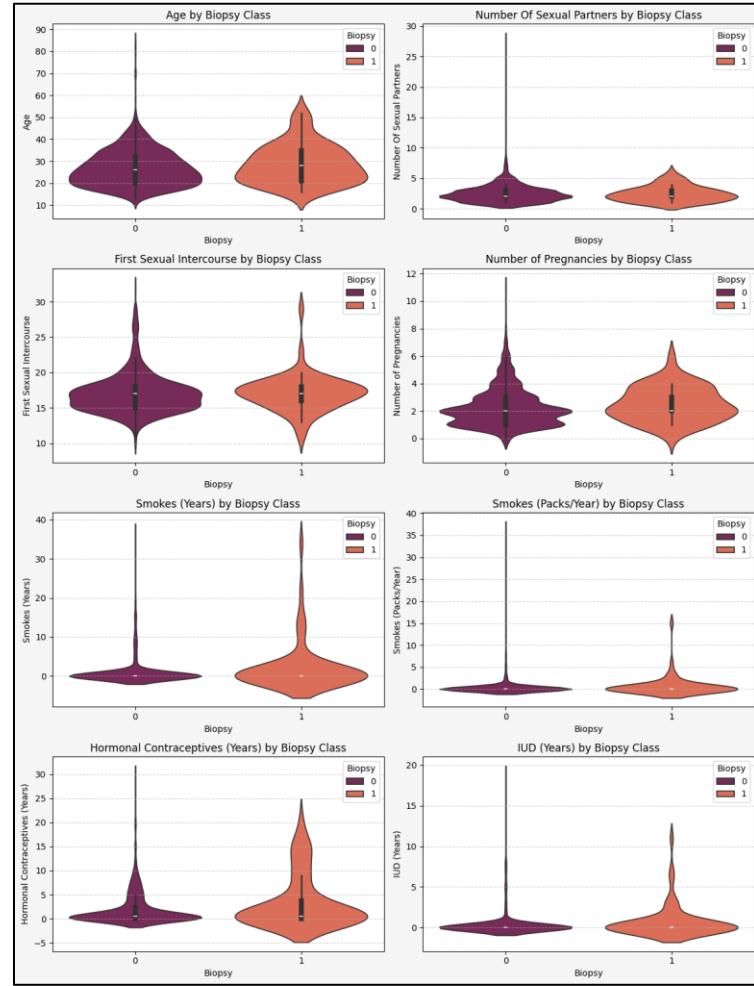
**Table 1:** VIF calculation display. We can see that there are features which show extreme VIF values. However, the decision drop or engineer them is not so simple, since they are all binary and connected to each other (part of the STDs lists). The best approach would be to save a list of the most severe cases and then try different experimentation approaches during the modeling process.

### 3.4. Feature Selection Outcome

The analysis described in this section led to the selection of the most statistically relevant and promising set of features, for each of the target variables. Results were used to generate a custom model-ready dataset for each target (*Table 2*).

## 4. Modeling Phase

This phase involved the development and evaluation of machine learning classifiers to predict four diagnostic outcomes related to cervical cancer screening. Each target - Biopsy, Schiller, Hinselmann, and Citology - was modeled independently using a subset of statistically significant features identified during exploratory analysis (described in previous sections). The modeling workflow emphasized proper data partitioning, class imbalance handling, consistent performance metrics, and comparative model evaluation. Three classifiers were tested for each target, with the aim of identifying the most effective approach for each prediction task, balancing accuracy, sensitivity, and interpretability.



**Figure 3:** Example violin plots of the distribution of the group (a) features per Biopsy target subclasses. There are some differences in value clustering, value spread and variances, many features do not seem to exhibit notable class differentiation.

<b>Biopsy</b>	Citology, Dx, Dx:CIN, Dx:Cancer, Dx:HPV, Hinselmann, Schiller, Stds, Stds (Number), Stds:Condylomatosis, Stds:Number Of Diagnosis, Stds:Vulvo-Perineal Condylomatosis
<b>Schiller</b>	Age, Biopsy, Citology, Dx, Dx:Cancer, Dx:HPV, Hinselmann, IUD, IUD (Years), Number Of Pregnancies, Stds, Stds (Number), Stds:Condylomatosis, Stds:Number Of Diagnosis, Stds:Vulvo-Perineal Condylomatosis
<b>Hinselmann</b>	Biopsy, Citology, Dx:Cancer, Dx:HPV, Schiller, Stds (Number), Stds:Number Of Diagnosis
<b>Citology</b>	Biopsy, Dx, Dx:Cancer, Dx:HPV, Hinselmann, Schiller

**Table 2:** Final feature set selection for each corresponding target variable.

#### 4.1. Model Pipeline Overview

The modeling pipeline was designed to maintain reproducibility, handle class imbalance, and rigorously evaluate model performance. A stratified 80/20 train-test split was applied to preserve class distributions in each target variable. Given the significant imbalance across all four targets (particularly Citology and Hinselmann), SMOTE (Synthetic Minority Over-sampling Technique), was applied to the training data to synthetically generate minority class samples. This ensured more balanced training without altering the test set for unbiased evaluation.

Model Performance was evaluated using a range of metrics, giving more weight to the best fitting ones for the imbalanced target situation. We examined the following metrics:

- **Accuracy:** Overall correctness of predictions
- **Precision:** Proportion of correct positive predictions
- **Recall:** Ability to identify actual positives
- **F1-Score:** Harmonic mean of precision and recall
- **ROC-AUC:** model's ability to discriminate between classes across thresholds

Visual diagnostics such as confusion matrices, ROC curves, and metric bar plots were used to further interpret model behavior.

#### 4.2. Models Tested

Three machine learning models were trained and evaluated for each target variable:

- **Logistic Regression:** a baseline linear model known for its interpretability and strong performance with smaller or imbalanced datasets (7).
- **Random Forest Classifier:** a robust ensemble method capable of handling non-linear relationships and interactions (8).
- **XGBoost Classifier:** a gradient-boosted decision tree ensemble optimized for performance (9). Two variants were tested: a default model, and a tuned model using GridSearchCV for hyperparameter optimization, with *recall* as the scoring function.

Other common classifiers such as Naive Bayes, Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), and Neural Networks were deliberately excluded for the following reasons:

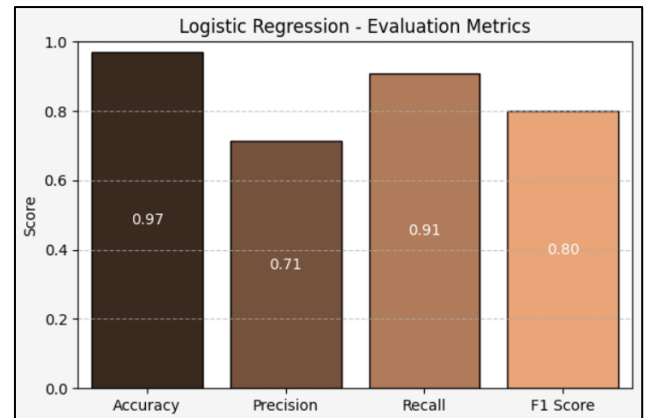
- **Naive Bayes** assumes conditional independence between features, which was deemed not suitable given the correlation between many of our clinical variables.
- **SVMs** and **K-NNs** are known to scale poorly with small or imbalanced datasets and can be computationally inefficient for grid-based tuning and probabilistic output.
- **Neural Networks**, while powerful, require larger datasets and more tuning to generalize effectively.

We believe that the chosen models achieve a balance between interpretability, robustness, and performance, making them well-suited to the clinical nature of the problem and the scale of the dataset. All models were trained using the same feature sets identified during EDA and were evaluated using identical metrics for fair comparison.

#### 4.3. Performance Analysis by Target

##### Biopsy

Logistic Regression achieved the best overall performance, with a high recall (0.91) and F1 score (0.80), making it well-suited for detecting true positive cases. It outperformed Random Forest and XGBoost in both discrimination and consistency (*Figures 5a, 5b and 5c*).



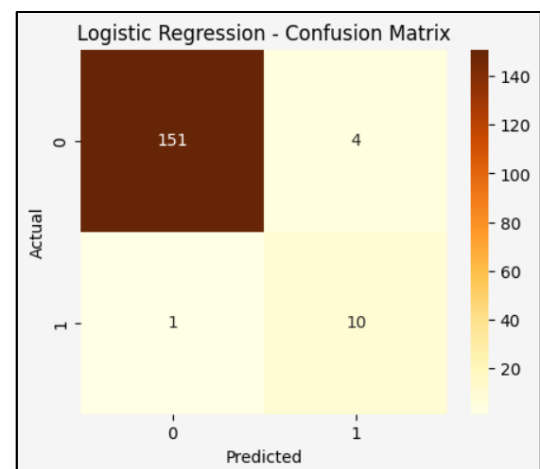
**Figure 5a:** Bar plot of the evaluation metrics for the Biopsy target (Logistic Regression model).

##### Hinselmann

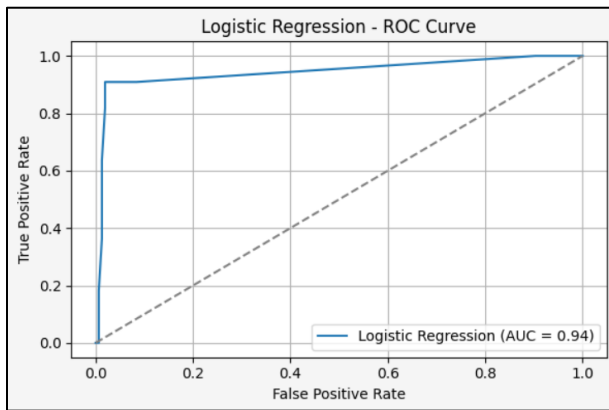
This target also favored Logistic Regression, which achieved the highest ROC-AUC (0.98) and recall (0.86). All models showed decent performance, but Logistic Regression's stability and interpretability made it the preferred choice (*Figure 6*).

##### Citology

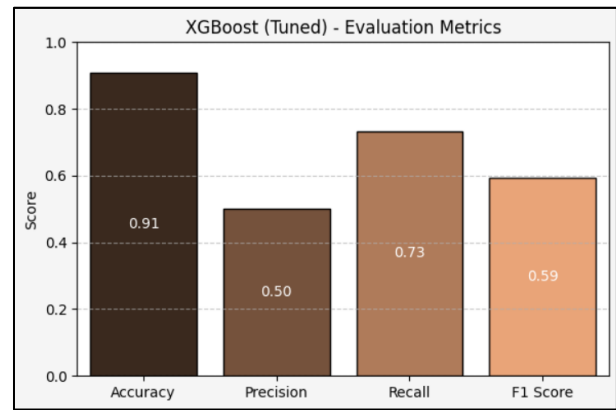
Due to extreme class imbalance, all models struggled to capture positive cases. Random Forest showed marginally better recall (0.33) and F1 score (0.32) compared to Logistic Regression and XGBoost, both of which had zero recall in some runs. Therefore, Random Forest was selected as the most suitable model despite the lower overall performance (*Figure 7*).



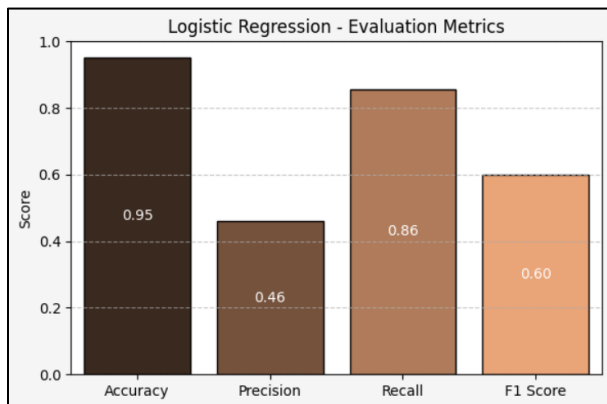
**Figure 5b:** Confusion Matrix for the Biopsy target (Logistic Regression model).



**Figure 5c:** ROC-AUC graph for the Biopsy target (Logistic Regression model).



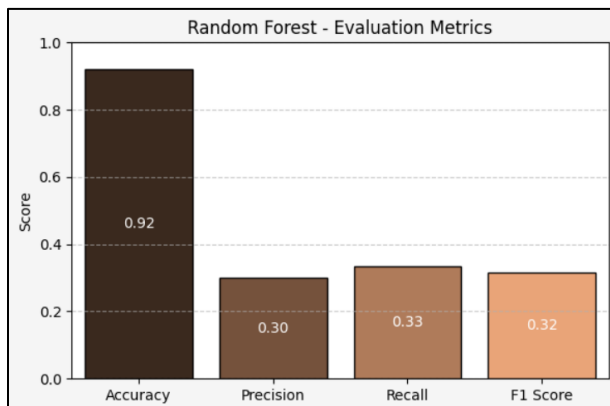
**Figure 8:** Bar plot of the evaluation metrics for the Schiller target (XGBoost model).



**Figure 6:** Bar plot of the evaluation metrics for the Hinselmann target (Logistic Regression model).

### Schiller

Logistic Regression again offered the most balanced results, with a recall of 0.67 and F1 score of 0.62. However, the XGBoost model achieved higher recall and top-level AUC at 0.91. If sensitivity and probabilistic generalization are our primary targets, then this is probably the most powerful model in this case (Figure 8).



**Figure 7:** Bar plot of the evaluation metrics for the Citology target (Random Forest model).

### Overfitting Checks

To assess model generalization and detect overfitting, performance metrics were computed and compared across both the training and test sets. Specifically, F1 Score, Precision, Recall, and Accuracy were plotted for each fold during training (via cross-validation) and then contrasted/compared against the same metrics on the test set. These comparisons were visualized using multiple line plots, to allow for an intuitive understanding of metric diversities between the two phases.

In general, the selected models exhibited *minimal overfitting*, as test performance remained closely aligned with cross-validated training metrics. Notably, Logistic Regression demonstrated the most stable generalization across all targets, while the other two main models (Random Forest and XGBoost) occasionally showed slight performance drops (particularly on imbalanced targets) due to their increased complexity and sensitivity to class imbalance. The use of the SMOTE algorithm helped mitigate this effect, resulting in balanced learning across classes during training.

### 4.4. Model Comparison Summary

The model outcomes across all four targets can be summarized (Table 3), with the best performing model for each task, along with the corresponding key-performance metric. While Logistic Regression consistently offered strong generalization and high recall for most targets, XGBoost proved more reliable for the Schiller target, and Random Forest also showed greater sensitivity in the most challenging case (Citology). Overall, the chosen models balance predictive accuracy with interpretability and robustness, highlighting their suitability for supporting diagnostic screening workflows in clinical settings.

## 5. Conclusion/Discussion

This project demonstrated the feasibility and effectiveness of applying machine learning techniques to predict cervical cancer diagnostic outcomes using patient health and behavioral/lifestyle data. Through a complete data science pipeline (from data cleaning and exploratory analysis to model selection and evaluation), robust classifiers were developed for four distinct targets. *Logistic Regression* / *XGBoost* emerged as reliable and interpretable baselines, particularly excelling across most targets, while *Random Forest* proved more adaptable in highly imbalanced scenarios. The findings highlight the potential for machine learning to assist in early screening and decision support in various clinical contexts.

Target	Selected Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Biopsy	Logistic Regression	0.97	0.71	0.91	0.8	0.94
Schiller	XGBoost (Tuned)	0.91	0.50	0.73	0.59	0.91
Hinselmann	Logistic Regression	0.95	0.46	0.86	0.6	0.98
Citology	Random Forest	0.92	0.3	0.33	0.32	0.65

**Table 3:** Final selected models for each target variable, and their performance on all corresponding metrics. It is worth noticing that even though the Accuracy metric is extremely high for all targets, in our case it is not really appropriate as a metric, due to the extreme target class imbalance. Recall is to be preferred, especially if we are interested in “catching” as many positive cases as possible, and also ROC-AUC, if we are interested in probabilistic generalizations.

## References

1. Cervical Screening [Internet].; 2025 [updated -04-10; cited Apr 14, 2025]. Available from: [https://en.wikipedia.org/w/index.php?title=Cervical\\_screening&oldid=1284909519#General\\_screening\\_procedure](https://en.wikipedia.org/w/index.php?title=Cervical_screening&oldid=1284909519#General_screening_procedure).
2. Fusco E, Padula F, Mancini E, Cavaliere A, Grubisic G. History of colposcopy: a brief biography of Hinselmann. Journal of prenatal medicine. 2008;2(2):19–23.
3. Schiller's test [Internet].; 2023 [updated -07-21; cited Apr 14, 2025]. Available from: [https://en.wikipedia.org/w/index.php?title=Schiller%27s\\_test&oldid=1166354287](https://en.wikipedia.org/w/index.php?title=Schiller%27s_test&oldid=1166354287).
4. SMOTE: synthetic minority over-sampling technique: Journal of Artificial Intelligence Research: Vol 16, No 1.
5. Fernandes K, Cardoso JS, Fernandes J, Salvador Sánchez J, Rodrigues JMF, Alexandre LA. Transfer Learning with Partial Observability Applied to Cervical Cancer Screening. In: Switzerland: Springer International Publishing AG; 2017. p. 243–50.
6. Supervised deep learning embeddings for the prediction of cervical cancer diagnosis [Internet].; 2018 [updated -05-14; cited Apr 14, 2025]. Available from: <https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+fact+ors>.
7. Peng CJ, Lee KL, Ingersoll GM. An Introduction to Logistic Regression Analysis and ReportingThe Journal of educational research (Washington, D.C.). 2002;96(1):3–14.
8. Garvey J. Random Forest Introduction. 2022 -01-25.
9. XGBoost: A Scalable Tree Boosting System. 2016 13 August.